Name: Yun Zhou (1009709442)
Assignment 4: INF2178 Technical Assignment 4

Study MRI data

Introduction

The dataset originates from a longitudinal study on MRI results of patients with/without dementia. The data includes multiple visits of participants, was categorized into groups based on dementia status. The study participants were marked by unique subject IDs with diverse backgrounds and ranging in age.

The data aiming to investigate the progression of dementia through MRI scans, it allows to explore whether lifestyle, genetic factors, and external treatment might impact the course of dementia. Therefore, search for answers that can transform lives.

1. **Data Cleaning and preparation**

The dataset has a total of 16 columns with 294 rows, presenting information across various dimensions of dementia research.

From initial observations, the dataset includes a mix of demographic information such as age, education, or F/M (Sex), as well as clinical assessment data (MMSE & CDR) and MRI result measurements data (eTIV, nWBV, ASF). Also, some missing data under SES and MMSE columns, there might some potential gaps in the data collection process and need further data cleanning.

To delve deeper, below are the research questions that this dataset can help answer:
Research Question 1: How does cognitive decline as measured by CDR, differ between groups diagnosed as demented, nondemented, and converted?

Research Question 2: How do changes in brain volume as measured by nWBV, differ between groups diagnosed as demented, nondemented, and converted?

Research Question 3: Whether the variances of cognitive decline and brain volume measures consistent across the groups?

The goal is to understand the correlation between the different factors (such as demographic groups or disease progression) and with cognitive decline and brain volume changes among the dementia-related groups.

2. **Exploratory data analysis**

First is to examine the data, including removing the “Hand” column as the value is constant across all subjects, imputing the missing values with the median under “SES” & “MMSE”, and transforming categorized variables “Group” & “M/F” into numerical format. As a result, there are no more missing values and encoding is complete after EDA with new data frame.

To visualize the dataset's distribution, histograms, boxplots and correlation matrix are created. From six **histograms** for continuous variables, the **age** distribution is nearly a normal distribution as well as "eTIV distribution", with age centered around 76, therefore having a middle to older-aged population as the risk of having dementia increases with age. Also, the physical brain structure is not directly impact by disease's cognitive nature. Per "Years of education distribution", the most common education levels are 12 years and 16 years, therefore a variety in educational background study participants. Per "**MMSE** Score distribution" is left-skewed, with large number of participants around score of 30, this means large number of participants had good cognitive function.

Per **Boxplot**, the CDR scores for the Nondemented group is concentrated at 0, whereas Demented Group has higher scores at 0.5 as expected and outliers. **Correlation matrix** output, there is strong positive correlation (0.89)between Visit and MR delay, the longer the delay the higher number of visits. The CDR score and MMSE has strong negative correlation (-0.7) so higher CDR scores are associated with lower MMSE scores. The eTIV and ASF has strong negative correlation (-0.99), whereas Age and nWBV has moderate negative correlation (-0.54) as brain volume decrease as age increase.

3. **Data analysis**

Research Question #1: How does cognitive decline as measured by CDR, differ between groups diagnosed as demented, nondemented, and converted?

Mixed AVOVA comparing 'Group' factor for dependent variable 'CDR':
The mixed ANOVA model 1 performs a mixed-design ANOVA on the "CDR" scores which as dependent variable, "group" as between-subject factor, "visit" is within-subject factor, and "subject ID" as random factor. Per results of the mixed ANOVA for 'CDR', p value for group is less than 0.01 (6.603e-46), therefore a highly statistically significant difference in CDR scores between the groups. P value for Visit is 2.51e-05 which is less than 0.01, therefore significant changes in CDR scores over the different visits. The interaction effect has p value of 5.73e-05, which is less than 0.01, meaning a significant interaction between the Group and Visit factors, thereby the changes in CDR scores due to visits are not constant across groups.

Overall, the group difference, change over time and interaction between groups and are statistically significant in terms of CDR scores.

Research Question 2: How do changes in brain volume as measured by nWBV, differ between groups diagnosed as demented, nondemented, and converted?

AVOVA for nWBV:
The mixed ANOVA model 2 is similar to the above model 1, but use "nWBV" as dependent variable instead of 'CDR'. The p value is 1.64e-03 and is less than 0.05 (significant level) therefore 0.01, indicating a significant difference in nWBV among the groups. The p value is also extremely low at 2.22e-17 for the visit factor, therefore a highly significant change in nWBV

over the visits. The p value is 2.19e-01 and over 0.05 (significant alpha level), therefore the effect of Visit on nWBV does not differ significantly among group.

To visualize the interaction, we use point plot to display the interaction between the 'Group' and 'Visit' factors for the 'CDR' measurements. per Interaction Plot, the demented group shows an increase in CDR scores whereas the Converted group's CDR shows dramatically increase on 2nd visit. However, there is little change in CDR scores for non-demented group.

Research Question 3: Whether the variances of cognitive decline and brain volume measures consistent across the groups?

Also, to perform Homogeneity of Variance test to check the assumption of variances across groups for 'CDR' and 'nWBV' dependent variable. For CDR, p value is 2.78e-09 and is less than 0.05 (significant apha level). The assumption is violated for CDR and the variances are significant. For nWBV, p value is 0.38 per levene's test which is greater than 0.05, therefore not statistically significant and the variance of nWBV are not different across three groups.

In summary, the result shows significant differences between the groups' CDR and nWBV scores and across visits. Also, the violation of homogeneity of variances for CDR might impact the validity of the ANOVA result.

Pairwise t-test post-hoc tests is conducted to further explore the differences between groups and across visits for CDR and nWBV.

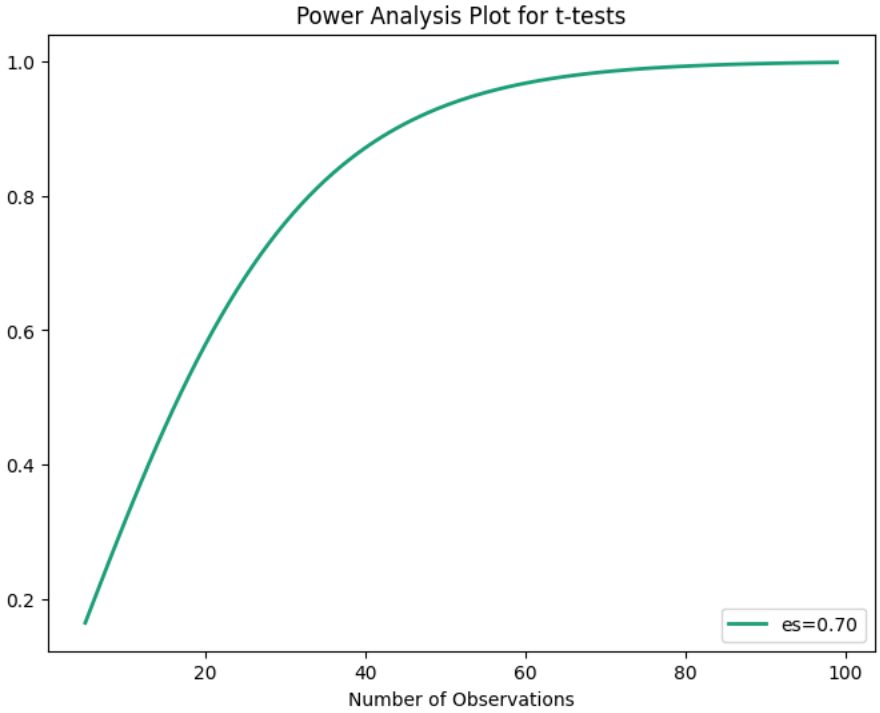
Pairwise post-hoc Tests for CDR: significant differences were observed between all pairwise group comparisons (Converted vs. Demented, Converted vs. Nondemented, Demented vs. Nondemented) with p value 0.000 for all 3 above, indicating strong evidence against the null hypothesis of no difference. However, for the interaction between group and visit, the Converted vs. Nondemented comparison at Visit 1 was not significant as p value is at 1 after correction for multiple comparisons, suggesting no significant change in CDR scores between these groups from Visit 1 to Visit 2.

Pairwise post-hoc Tests for nWBV: In contrast, for nWBV, there is significant difference between the Demented and Nondemented groups as p value is 0.001(less than 0.05), while other comparisons not showing significance after adjusting for multiple comparisons. Overall, there is significant differences in nWBV scores observed from Visit 1 to Visit 2 for the Demented vs. Nondemented groups, highlighting the nuanced changes over time in brain volume between these specific groups.

4. **Statistical power analysis for t-tests** (power =0.91, alpha =0.05, effect size =0.7)

We conduct a power analysis for an independent sample t-test to determine the sample size needed to have a 91% chance of detecting an effect size at 0.7 if it actually exists, while

maintaining 5% of chance of falsely detecting an effect if there is none. And also to show how the power of the test increases with the number of participants.



Per the plot output, the calculated sample size needed per group is 45.05 to achieve a power of 0.91 for detecting and effect size of 0.7 with alpha of 0.05. therefore, we would need at least 46 participants in each group for the study. The point where reaches the top shows that increasing sample size further will result in diminishing returns even increasing power of test.

Overall, power analysis result is to have at least 46 per group, both the demented and nondemented groups meet and exceed this requirement. However, the converted group does not meet with having 14 participants. Therefore, demented and nondemented group is sufficiently powered.